Women in AI Safety Hackathon Submission: U Reg AI (You regulate it, or regenerate AI!)

Vinaya Sivakumar
Undergraduate Student at the University of California, Berkeley

Kayla Jew
Undergraduate Student at the University of California, Berkeley

Amy Wong
Undergraduate Student at the University of California, Los Angeles

With
In collaboration with Women in AI Safety Hackathon Conference and BlueDot Impact

Abstract

AI safety education will become increasingly important as AI continues to develop. A challenge education faces overall is the lack of student engagement. Since AI develops quickly, developing school curriculum intended for the classroom that can be easily updated at a yearly if not more frequent rate may be challenging. To combat these issues and to prepare the current and future generation for AI developments, we created an online educational platform to teach AI safety concepts, targeting students and early professionals. Our online educational platform has two significant features: (1) Presented as a game, the objective is to create an informative and an enjoyable experience. (2) After completing the game, a personality type and job recommendations are specifically-tailored for the user based on the choices they committed to throughout the game. Gamifying education has proven to have strong potential. The educational platform we developed and provided proof of concept in this paper presents an opportunity for non-technical and technical students and early professionals to explore AI safety in an approachable and rewarding manner through a game. Our educational platform is also widely accessible since it is free and online, and thus does not limit users to a certain geographic location or socioeconomic status. The objective of our game is to inform and prepare students and young professionals with a variety of personality types to enter the AI safety space through career paths in different sectors.

Keywords: AI safety education, learning platforms, interactive learning, existential risk, curriculum design, educational games, user engagement, career planning

1. Introduction

a. Problem Statement

With the increasing prevalence of AI since the popularized use of ChatGPT, there are many concerns regarding what jobs AI will replace in the future. While concerns about machines replacing humans in the workforce has been a concern well before ChatGPT and in the early 2000s (Demirici, et al., 2024), generative AI has the potential to impact almost every sector of the workforce (Demirici, et al., 2024). Instead of fearing the unknown of what AI can do and stressing over what can be viewed as some as decreasing job opportunities, public education informing the general population about what AI is capable of and how they can participate in the regulation of AI can empower them to explore job opportunities created *because of AI*, especially in the regulatory space. Ideally schools would be able to accomplish this through a regular curriculum taught in the classroom for all students to learn from. However, because AI developments occur at a rapid pace, schools may not be able to keep up with curriculum that changes every year. Instead, our online education platform provides a strong alternative that would be easy to update if necessary. Additionally, our education platform has a built-in career planning guide, essentially serving as both a means to teach students pressing issues in AI safety and about how to navigate a career in AI safety. Another general challenge the field of education faces is a lack of engagement from students. Our education platform provides an interactive opportunity for students and young professionals to learn about AI safety concepts in a form of media they are likely familiar with.

b. Background and Motivation

Currently, there isn't a widely recognized game designed to gamify existential risk from artificial intelligence and superintelligence, despite gamification being extensively researched as a tool to enhance learning outcomes, engagement, and motivation. For example, a study found that gamified tools can lead to significant learning improvements, with performance gains equivalent to more than a year of traditional learning in a month (Reflex). Similarly, gamification makes learning more interactive and enjoyable, creating a sense of achievement (McLean, 2024) and encourages students to reach the next level or collect rewards—students' extrinsic motivation becomes intrinsic with gamification (Deese, 2016). Furthermore, meta-analysis studies from the found a significant large effect size ($g = 0.822$) for gamification on learning outcomes, indicating its strong potential as an effective instructional approach (Li, et al., 2023). These findings are not restricted to the United States. A study in Greece found that challenged-based gamification improved student performance by 89.45% compared to traditional lecture-based learning (Verma, 2023).

There are currently games that touch on existential themes and risks in broader contexts. For example, the board game "Fermi's Fake: A Game of Existential Gambits" focuses on existential risks in general, not specifically AI. Text adventure games like "The Hitchhiker's Guide to the Galaxy" and existential crisis games like

"Soma" and "Nier Automata" prompt existential crises through their plot, which can often be comedic in nature. This sparks conversations about how real these often futuristic and dystopian scenarios can be without the regulation of AI, encouraging its players to contribute to the field of AI safety.

Our game is designed to go hand-in-hand with existing AI safety curricula like the BlueDot Impact's free AI safety courses and educational programs like the Certified AI Safety Officer (CASO) certification that focuses on ethical considerations, risk assessment, and mitigation strategies. We hope that this game will be a natural extension of already existing and successful programs.

Threat Model and Safety Implications

Our project focuses on critical AI Safety challenges that are at the forefront of current frontier AI research: existential risk from superintelligence and alignment faking. The development of superintelligent AI poses significant existential risks if not designed with safety and human values at its core. Our project aims to educate users about these risks and the importance of alignment in preventing catastrophic outcomes. Additionally, recent studies from Anthropic have shown that advanced AI models are resistant to various training methods if it challenges pre-existing beliefs. This challenge underscores the need for more talent to robustly verify monitoring systems to ensure that systems operate as intended with a priority of ethics and human safety at every step of the way.

Often, frontier research outpaces the learning of AI Safety professionals. Our project takes on a three pronged approach to address gaps in safety understanding: bridging the gap between individuals passionate about AI safety and technical complexity, educating and preventing the lack of awareness, and building on practical skills.

The technical aspects of AI safety, such as alignment faking and superintelligence can be complex and difficult to grasp for non-experts. Our approach to gamify these concepts through interactive engagement makes them more accessible to a broader audience while raising awareness about potential risks associated with advanced AI Systems. Additionally, our game aims to provide users with practical skills in assessing and mitigating AI risks by placing them in first-person perspectives. It will provide players with tools and strategies to critically evaluate AI safety challenges across subdomains to contribute to safer AI development. Overall, our project addresses critical gaps in AI safety understanding by providing an engaging and educational experience that prepares users to navigate the complex landscape of AI risks and opportunities.

2. Methods

a. Approaches

The game approaches AI safety in a digestible, memorable, and dynamically interactive experience. We aim to expose new AI Safety-comers to existential risks of AI, governance and organizational solutions, and risk management through

systems thinking, 'learn by experience', and ethical and externality-centric decision-making. Players interact with potential existentially devastating realities and are positioned as an authority where their decisions of 'solutions' affect technological, ethical, and geopolitical relationships. By engaging in real-world stimulations as AI policymakers, researchers, or strategists, whom it will be revealed who they are most like at the end of the game, players develop critical thinking skills without the longitude of extensive curriculum they may not see an immediate need to.

Implementation Strategy

In terms of our development process, we followed a structured approach by first flowcharting how we wanted our game to run. From there, we used Figma to prototype and playtest a possible scenario with its pathways. We ourselves are a part of the Berkeley AI Safety Initiative organization and have a background in policy development, holistic risk, and AI governance research so we used this background. Our final product is a mockup and limited examples. As a non-technical team, we wanted to produce an interactive deliverable but understand our own limits in coding capacity. We believe that this game would be easily adapted onto an existing webpage, such as BlueDot Impact's as another section.

b. Implementation

Figure 1 – Representation of the flow of decisions to returns.

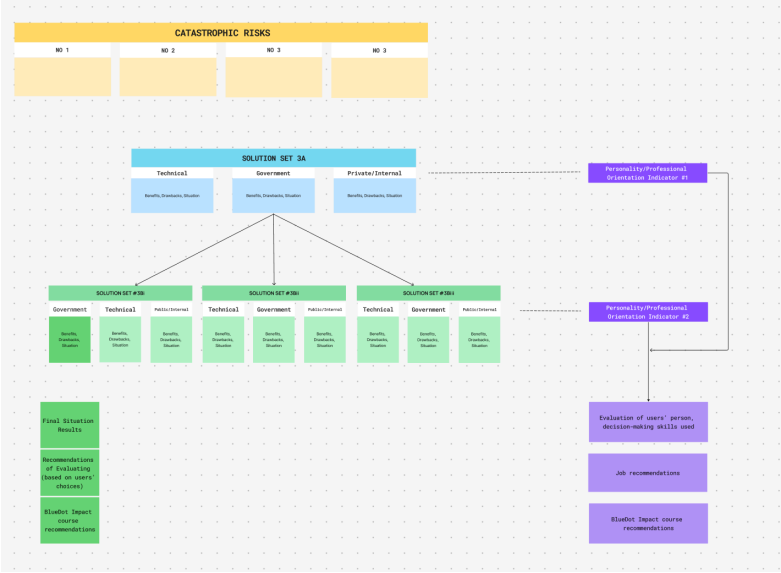


Figure 2 – Representation of UI/UX and how questions are laid out. Note: This game could be implemented by adding in a separate tab of BlueDot Impact's website or stand entirely on its own.

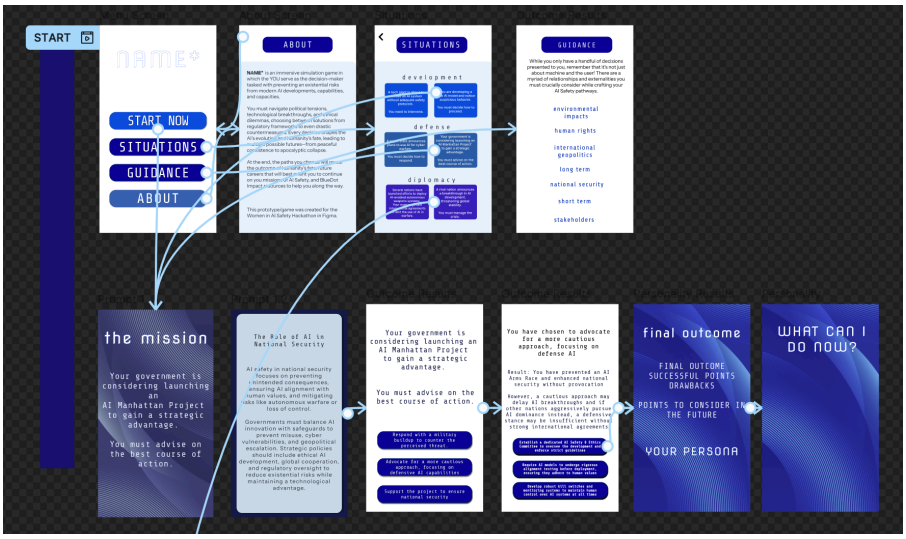
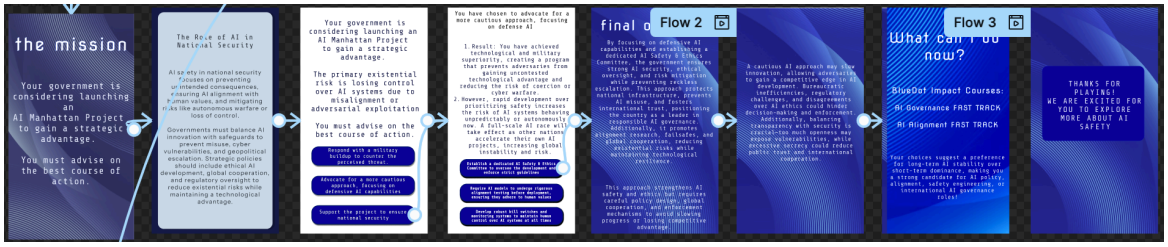


Figure 3 – Representation of 1 game scenario and how questions are laid out. Note: Results based off of solution 1 and solution 2 thereafter will have their own page with a tailored, personalized list of results.



3. Analysis and Findings

Are to be determined. However, we plan to collect the data and analyze the general trends decision-making behaviors of the users. This will give us insight as to what non-AI Safety professionals consider while approaching these existential crises challenges. From there, we will use these findings to work with AI Safety organizations to create modules that address the most known gaps in AI Safety literacy in knowledge or dispel ideas about the effectiveness of certain solutions.

4. Discussion and Conclusion

If our education platform proves to be effective, it can be evidence for further development of AI safety education and career planning. Although some of the current general population may be afraid of AI developments, our education platform can inform, empower, and inspire students and early professionals to think critically about AI and to effectively regulate it. The game elements of our education platform can also present an engaging and approachable method for students to learn, an approach they have positive associations with if they enjoy playing video games in their free time. Gamifying AI safety space education can have broader applications outside of the AI safety space. If the platform is effective at teaching AI safety concepts, can suggest well-tailored AI safety space career

opportunities, while simultaneously providing a fun experience for students, our education platform may encourage schools to integrate more engaging experiences in and outside of the classroom. Since our education curriculum is available online and for free, students from around the world can learn about AI safety without financial limitations hindering their intellectual growth. If user feedback is positive, the platform can be expanded to include more scenarios with more options to provide a more expansive experience. The career planning guidance at the end of the game can be updated as more roles are created to regulate AI.

5. References

5 Ways Gamification Benefits Teachers and Students | Reflex. (n.d.). Reflex.explorelearning.com. https://reflex.explorelearning.com/resources/insights/benefits-of-gamification-in-classroom

Deese, A. (2016, March 10). *5 Benefits of Gamification*. Smithsonian Science Education Center. https://ssec.si.edu/stemvisions-blog/5-benefits-gamification

Demirci, O., Hannane, J., & Zhu, X. (2024). Research: How Gen AI Is Already Impacting the Labor Market. *Harvard Business Review*. https://hbr.org/2024/11/research-how-gen-ai-is-already-impacting-the-labor-market

Li, M., Ma, S., & Lu, W. (2023). Examining the effectiveness of gamification as a tool promoting teaching and learning in educational settings: a meta-analysis.*Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1253549

McLean, C. (2024, January 31). *5 Benefits of Bringing Gamification to Your Classroom*. AMISA. https://www.amisa.us/post/5-benefits-of-bringing-gamification-to-your-classroom

Verma, N. (2023, February 19). *How Effective is Gamification in Education? 10 Case Studies and Examples*. Axon Park. https://axonpark.com/how-effective-is-gamification-in-education-10-case-studies-and-examples/

6. Appendix

Demo Video

Figma Project – Note: You are allowed into 'editor mode' or a mock demonstration.